

The exam will consist mostly of multiple-choice questions. Maybe some simple computations will be required, so keep your phone or calculator close by.

Regression Topics:

- Multivariate Linear Regression
- Polynomial Regression
- Support Vector Regression
- Decision Tree & Random Forest Regressions

The questions will be mostly to see how well you understood the intuition and how the models work. For example a possible question could be about how random forest attributes the prediction values once that the splits are made (answer: by taking the average of all y values of the data points in the split). Or how it makes the splits, i.e. what metrics it minimizes. Same could be asked about the linear regression or SVR. While I will not ask you to write out the formulas of the functions that are minimized (also called cost functions), I could ask you for example to compute the predicted value for some split of a decision tree (you will not need to use any calculus and solve any optimization problem, but you will need to understand what steps the model makes in order to find the predictions).

Classification Topics:

- Logistic Regression
- Decision Trees & Random Forest Classification Models
- K-NN
- Naïve Bayes Model
- Support Vector Machines

Similarly the questions will be mostly about the intuition of the models, that means you have to understand how the models work, how they split data into classes and how they assign new data points to predicted classes. There will also be questions about the pros and cons of these models (you can find everything in the slides).

Time Series Models:

- What is special about time series (trend seasonality, etc.)
- Stationarity: the idea of stationarity (what is it), how to test for it (DF and ADF tests), how to correct for non-stationarity and why is that necessary.
- What are the different approaches to model time series (ML and Statistical models).
- The AR(p) model – intuition and how the number of lags are selected (PACF)
- The MA(q) model – intuition and how the number of lags are selected (ACF)
- The ARMA(p,q) model

- The ARIMA (p,d,q) model – intuition, what does ARIMA add to our modeling capabilities in comparison to ARMA? What are the two approaches to selecting d (remember Auto-ARIMA and DF test)?

Other topics

K-fold cross validation method

Scaling of Variables